**Segmenting US Cities based on similarity**

khuram k

September 5th 2019

## Introduction 1:

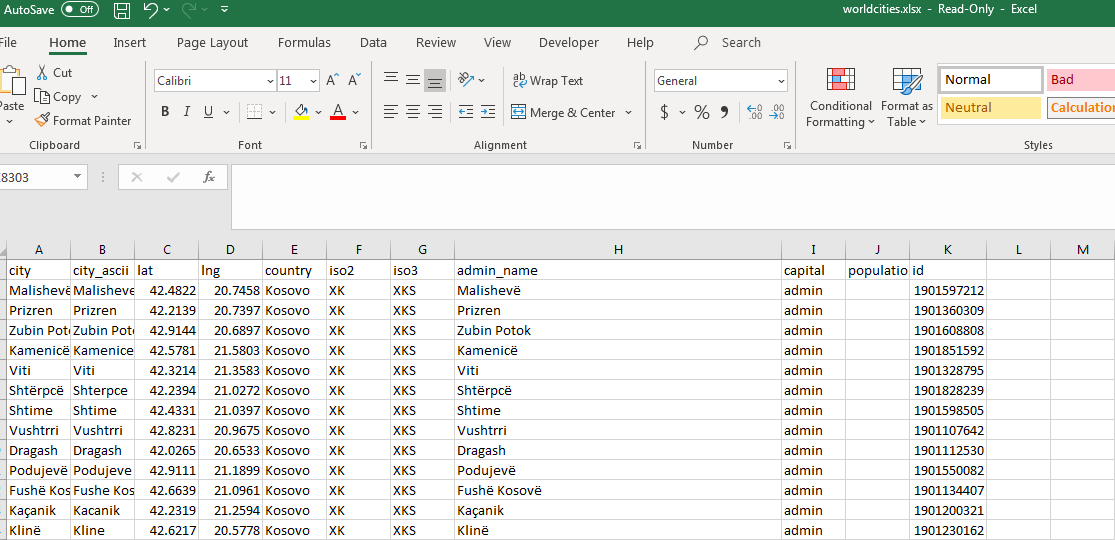
We have seen various ways cities are categorize based on population density wise, by weather conditions and various other ways. But have you ever taught of looking if I am living in city and planning to move to another city and want to know how similar to other city in terms of accessibility and shopping centers, schools, colleges, health care and other terms. These are something which everyone like to know before considering for any job moment and planning to move with family. So to tackle this and understand US cities how they are similar to another city, I have build the machine learning model that can cluster all US cities into 5 segment , putting each and every cities into this segment/cluster based on their similarities.

## Machine Learning Model used:

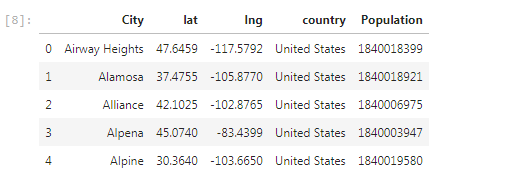
I used Machine learning unsupervised K-Mean Clustering model that helps classify all US Cities into cluster of 5 and segment the cities into 5 cluster based on their similarities to one another.

## Data Acquisition and Cleaning

I used free source available data on world cities and imported to dataframe from CSV. The data we were interested was all US cities and latitude and longitude. The initial raw data looks like

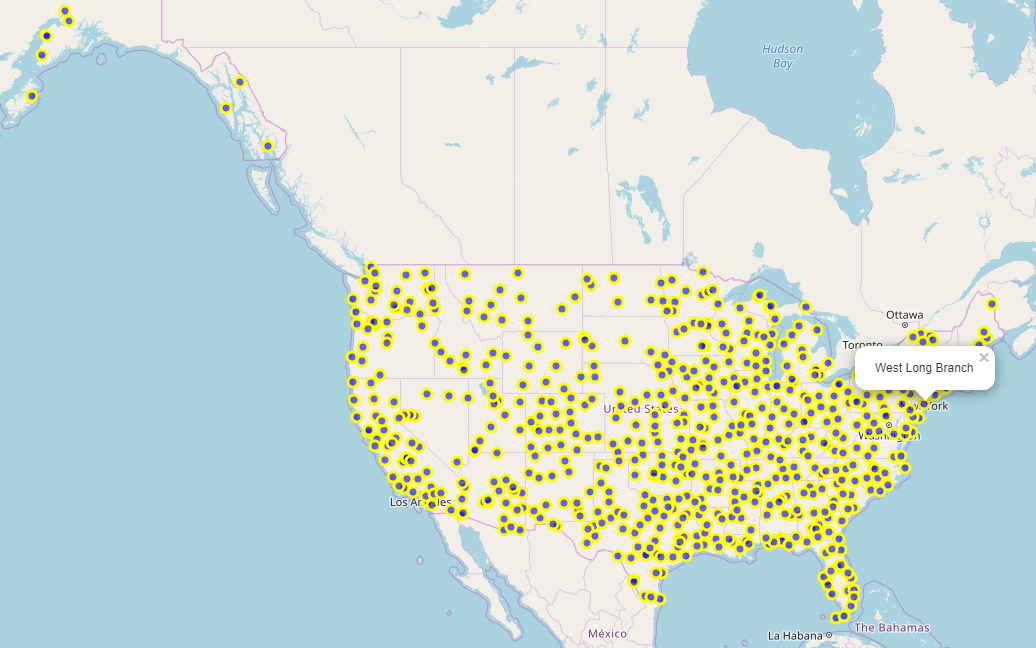


I imported these data into dataframe and extracted only US cities and kept only the important columns which are latitude and longitude as this was needed for my API Calls to foursquare to extract all US Cities to get venues details for machine learning modeling. There’re around more than 4000+ US cities in my dataframe , due to limitation of my Jupyter notebook processing for Map visualizations and also limited access to foursquare API (as I can make only 2500 calls per day). I had to limit the data in my dataframe for which I filtered out cities with similar name keeping the cities from duplicate names with highest populations and also eliminating cities which are close to one another by removing cities with same latitude and longitude (with first 2 digits). This resulted in shorter list of 527 cities to work easily with my developer/limited access. Once after data cleaning, the cleaned and ready to machine learning model dataframe looks like



## Data Visualization before cluster

Ext the clean data was plotted on map to visualize to see the data distribution across US map to get the clean idea of Cities selected using circular marker



As we can see the data/Cities we selected are well corners of the US city to make sure when we cluster it and if you are planning to go to any cities cluse to cities in map, you can always evaluate looking at the cities close to the one available in maps.

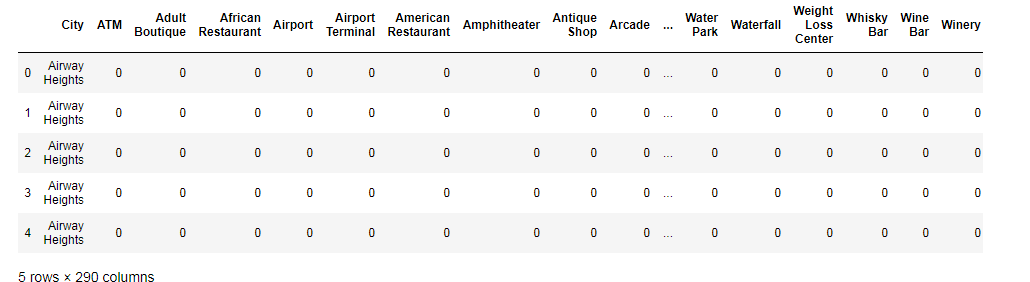
## Data Normalization:

After the API Calls to Foursquare that venues details were retrieved for all the selected cities and data we retrieved is categorical (Venue category ) which is our interest for building model ,

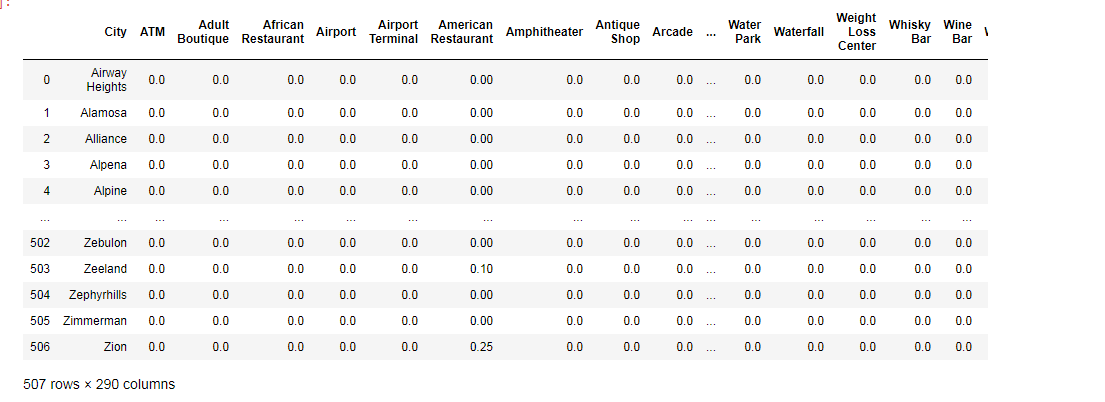
Dataframe post foursquare API call and appended the Venue details (last column) to my original cleaned dataframe (that has US cities name, population and city latitude and longitude)



As I was using Sklearn package for building my model, I had to transform my categorical values for column ‘Venue Category’ into numerical using one hot coding and data was transform into numerical



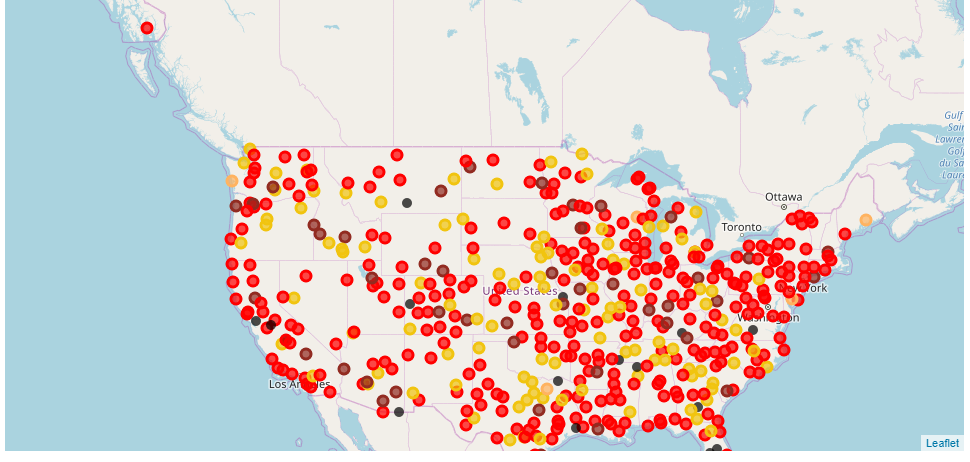
and finally normalized data by grouping on cities so that we can start building machine learning. This is the data we will depend on building and predicating rest of the modeling and review



## Clustering – K Mean Technique

I applied the machine learning model unsupervised technique of K Mean clustering with 5 segment to keep it simple and straight when segmenting and grouping the cities into 5 cluster based on similarities to make it easy for people to look into this clustering and able to take a decision if their city is close to which other city of their interest.

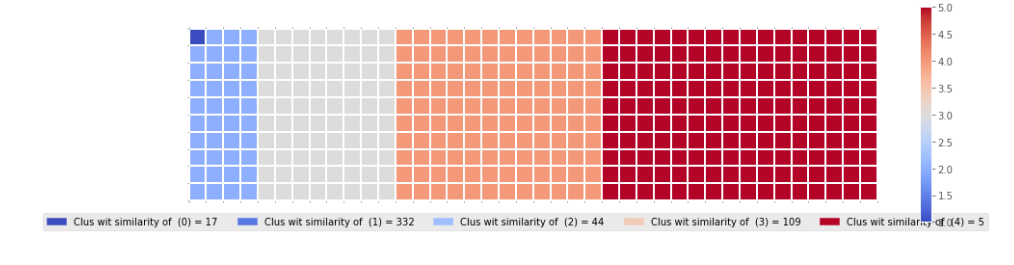
After predication, the clustered US cities were mapped to US Map to show in 5 different color (each for each segement to color code) to show the similarities as shown below.



If you hover are click any of the dot it display the city naes and color coding of circular marker indicate the cluster of the cities into five cluster

## Predication /Conclusion:

To further evaluate distribution of this cities based on alike-ness, a waffle chart was created to see the percentage or distribution of US cities based on similarities



It can predicated or conclude that

* + 65% of US cities are alike to each other falls in cluster 1,
  + 22% under another cluster likewise and
  + Only 1 percentage which is cluster (blue) around 5 cities tends to stand out and look to be unique and very dissimilar and look for it.

## Future Enhancement to model:

As this model is just an start, a lot of feature can be added and enhanced such as

* Adding details for each of cluster 10 – 15 most common point of interest
* Building visualization of cluster cities with embedded names
* Predicating or testing other cities of the world, if we like to know example London is close to which city of US, are just few usage of this model